Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 62 (2015) 221 – 227

Procedia Computer Science

The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)

A Review of the Advances in Cyber Security Benchmark Datasets for Evaluating Data-Driven based Intrusion Detection Systems

Adamu I. Abubakar[a], Haruna Chiroma[b],*, Sanah Abdullahi Muaz[c], Libabatu Baballe Ila[d]

[a]*International Islamic University Malaysia, Faculty of Information and Communication Technology, Kuala Lumpur, Malaysia*
[b]*Federal College of Education (Technical), Department of Computer Science, Gombe, Nigeria*
[b]*University of Malaya, Department of Artificial Intelligence, Kuala Lumpur, Malaysia*
[c]*Bayero University Kano, Faculty of Computer Science and Information Technology, Department of Software Engineering, Kano, Nigeria*
[d]*Bayero University Kano, Faculty of Engineering, Department of Machanical Engineering, Kano, Nigeria*

Abstract

Cybercrime has led to the loss of billions of dollars, the malfunctioning of computer systems, the destruction of critical information, the compromising of network integrity and confidentiality, etc. In view of these crimes committed on a daily basis, the security of the computer systems has become imperative to minimize and possibly avoid the impact of cybercrimes. In this paper, we review recent advances in the use of cyber security benchmark datasets for the evaluation of machine learning and data mining-based intrusion detection systems. It was found that the state-of-the-art cyber security benchmark datasets KDD and UNM are no longer reliable, because their datasets cannot meet the expectations of current advances in computer technology. As a result, a new ADFA Linux (ADFA-LD) cyber security benchmark dataset for the evaluation of machine learning and data mining-based intrusion detection systems was proposed in 2013 to meet the current significant advances in computer technology. ADFA-LD requires improvement in terms of full descriptions of its attributes. This review can be used by the research community as a basis for abandoning the previous state-of-the-art cyber security benchmark datasets and starting to use the newly introduced benchmark dataset for effective and robust evaluation of machine learning and data mining-based intrusion detection system.

Peer-review under responsibility of organizing committee of The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)

* Corresponding author. Tel.: +60142256283.
E-mail address: adamu@iium.edu.my

Peer-review under responsibility of organizing committee of The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)
doi:10.1016/j.procs.2015.08.443

Keywords: Cyber Security; Benchmark datasets; Cyber Crimes; KDD; ADFA Linux dataset; UNM

1. Introduction

Network attacks are malicious activities intended to disrupt, deny, degrade or destroy vital information and services residing in the host computer network. The attacks on computer networks are performed through data streaming on network computers with the intention of compromising the computer system's integrity, confidentiality, or the computer network system's availability [1]. Cyber crime has significantly contributed to the loss of billions of dollars in recent times, with a potential negative impact on the global economy if the trend continues without preventive measures [2-3]. The advent of wireless access has changed the pattern of cyber crime, and perpetrators who can afford to subscribe to the service are now using wireless access to perpetrate cyber crimes comfortably from home, without even going to Internet cafes [4]. The penetration of modern wireless systems is computationally expensive for the hacker, but is by no means impossible [5]. It has been acknowledged that the increasing concern about the penetration of wireless networks remains a threat to modern cyberspace [6].

It can therefore be deduced that the security of computer systems is imperative in modern life. Security is of paramount importance in all types of system, regardless of the type of computer, ranging from standalones operated by an individual user to computers in large corporate networks. Research is increasingly focused on critical wireless information conduits, and it is imperative to ensure maximum and effective security of these networks [6]. It is necessary to put into place stringent security measures to minimize and possibly avoid the intrusion of cyber criminals. Cyber crimes require a global approach, because the Internet and other networks have no geographical boundaries; once an attacker has access to wireless connections at any location, he can commit cyber crime anywhere in the world. To avoid cyber attacks and minimize their impact, detection traced through a host system's calls has been an active research area for several decades [7]. Recently, there has been an unprecedented increase in the number of systems or algorithms for the detection of cyber attacks on a host, with a low rate of false alarms and high levels of accuracy in detecting anomalies [8]. The intrusion detection system protects the computer from unauthorized accessed [9].

In the literature, several machine learning [10] and data mining techniques [11-12] were used to propose intrusion detection systems. Typically, newly proposed and existing intrusion detection systems' performance is evaluated using cyber security benchmark datasets. For example, genetic algorithms and support vector machines were hybridized to create intrusion detection systems. The hybrid of the genetic algorithm and support vector machine was evaluated using the KDD Cup 1999 dataset [13]. Moradi and Zulkernine [14] used neural networks to design an intrusion detection system which was evaluated using this benchmark dataset. Helali [15] reviewed the application of data mining techniques to create intrusion detection systems, finding that researchers relied heavily on the KDD and Computer Science Department, University of New Mexico (UMN) cyber security benchmark datasets [15]. However, the KDD and UMN benchmark datasets are no longer relevant to the modern computer age because of significant advances in computer technology [6].

The optimal solution for intrusion detection has not yet been found. However, there are significant advances in improving the performance of existing cyber security systems [15], and studies are expected to flood the literature in the future search for optimal solutions.

In this paper, in view of the importance of cyber security benchmark datasets in evaluating intrusion detection systems before their real-life application, we review advances in order to present an appropriate updated benchmark dataset, relevant to current computer technology.

The rest of the paper is organized as follows. Section II reviews recent reports attempting to bridge the gap between cyber security and cyber crime. Section III reviews advances made on cyber security benchmark datasets. Section IV presents the studies that used current cyber security benchmark datasets for evaluating intrusion detection system, before the concluding remarks are presented in Section V.

2. Recent Reports on Cyber Crimes

The Internet Crime Complaint Center (IC3) established by NW3C/BJA and FBI-United States of America with the intention of fighting cyber crime, released its report on cyber crime in 2013. The IC3 was established on May 8 2000, receiving the complaints of cyber crime victims across the world. Figure 1 shows the steep increase in the number of complaints, reaching its maximum in 2009 before starting to drop. The IC3 encourages public awareness of how to establish immunity against cyber criminals. In 2013, it received 262,813 (see Figure 1) complaints, with a total dollar loss of $781,841,611 (Table 1), which represents a 48.8 % increase in reported losses over 2012 ($581,441,110).

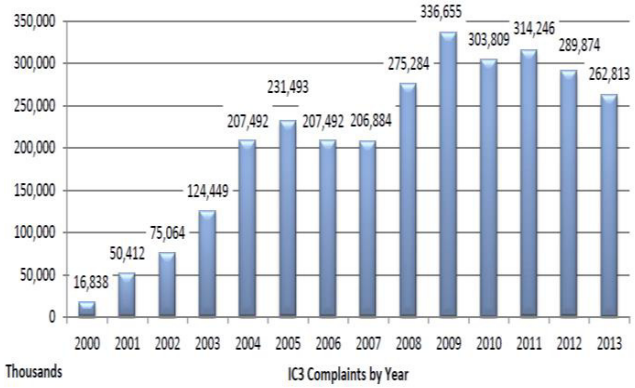

Fig. 1. The number of IC3 complaints received from 2000 to 2013. Source: [16]

Table 1 present demography of the conplainants in 2013, by males, females and combined, with the value of their lossess.

Table 1. The 2013 demography of the complainants including total losses in dollars [16]

Age Range	Male Count	Male Loss	Female Count	Female Loss	Total Complaints	Total Combined Losses
Under 20	5,194	$103,298,649	3,602	$2,364,515	8,796	$105,663,164
20 – 29	24,549	$42,144,452	23,483	$23,619,502	48,032	$65,763,954
30 – 39	28,391	$71,022,425	26,389	$41,784,048	54,780	$112,806,473
40 – 49	26,668	$89,559,205	29,170	$70,355,407	55,838	$159,914,612
50 – 59	29,220	$93,705,383	26,239	$83,858,340	55,459	$177,563,723
Over 60	23,074	$87,244,816	16,834	$72,884,870	39,908	$160,129,686
Totals	137,096	$486,974,929	125,717	$294,866,681	262,813	$781,841,611

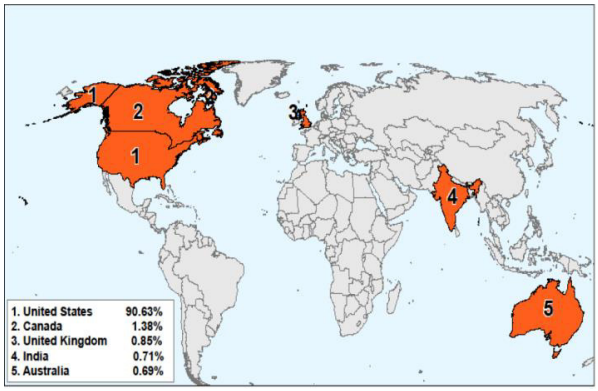

Fig. 2. The top five countries ranked by the number of complaints received. Source: [16]

The IC3 continues with its efforts to inform the public about the crimes perpetrated by cyber criminals, publishing service announcements and providing useful tips to Internet users. The successes recorded by the IC3 have attracted the attention of the international community. As a result, countries like Canada, the United Kingdom and Germany use it as a model for setting up similar centres against cyber crime. In a continuous effort to support foreign law enforcement, the IC3 is involved in the preparation of statistical reports of cyber crimes for specific countries. These reports are disseminated to hundreds of complaint referrals through FBI legal attaché offices throughout the world. The map shown in Figure 2 indicates the first five countries ranked according to the number of cyber crimes victim complaints reported in 2013 [16]. This clearly shows that cyber crimes have no geographical boundaries, and require a global approach as discussed in section I.

With appropriate and effective cyber security measures in place, those billions of dollar losses from cyber crime can be safe and used for further developing the world economy. Cyber crimes could potentially bankrupt legitimate businesses,and thus increase the rate of unemployment in society.

3. Cyber Security Benchmark Datasets

KDD was the pioneering benchmark dataset for evaluating a newly proposed intrusion detection system or algorithm. It was created by the Defense Advanced Research Projects Agency (DARPA) and other interested institutions to provide the benchmark dataset, because at that time there was no standard for the evaluation of a proposed intrusion detection system. This standard cyber security dataset attracted the attention of the research community, who started to use it as a benchmark for the evaluation of intrusion detection systems. The system used for the generation of the benchmark dataset was a Solari-based system, which can be easily integrated and monitored [6].

The dynamic nature of computer technology and revolutions in the computer industry due to rapid technological advancement started to place KDD at a disadvantage because its dataset could no longer provide the required accuracy for evaluation, and its results were no longer generally acceptable in the research community [17]. As a result, several cyber security benchmark datasets were developed to cope with the dynamic nature of computer technology, for example, the UNM [18] benchmark dataset proposed in 2004. However, this was criticized for failing to meet current trends in computer technology. UNM's limitations, in that it did not actually sample the targeted operating system and its scope was limited, meant that it did not succeed in replacing KDD The problem with the [6]. The KDD dataset continued to flourish in the research community, mainly due to the lack of a competent alternative [19].

The need to provide a reliable alternative cyber security benchmark dataset motivated Creech and Hu [6] to generate an alternative dataset that matched modern computer technology and the corresponding developments in cyber crime.. Using the old datasets to evaluate new intrusion detection systems would produce inaccurate and misleading results.

Creech and Hu [6] proposed the ADFA Linux (ADFA-LD) cyber security benchmarks datasets for evaluation of intrusion detection system. The host operating system for the generation of ADFA-LD was Ubuntu Linux version 11.04, which is one of the operating systems used across the globe. Its configuration offers various functions including

the sharing of files, database, remote access, as well as a web server. The Ubuntu 11.04 has Linux kernel 2.6.38 which is fully patched. FTP, SSH and MySQL 14.14 are enabled, based on default ports. Apache 2.2.17 and PHP 5.3.5 are installed for providing web-based services. Furthermore, TikiWiki 8.1 was installed as a collaborative web tool. The data structure for ADFA-LD is as follows: normal training data has 833 traces, normal validation data has 4,373 traces and attack data has 10 attacks/vector. The proposed ADFA-LD cyber security benchmark dataset has a closer resemblance between the attack and normal dataset, unlike the KDD cyber security dataset. In addition, the ADFA-LD actually represents the updated and modern cyber attacks of current situations. The ADFA-LD is freely available online and can be found in [20]. Table II presents the major cyber security benchmark datasets.

Table 2. Generations of the major cyber security benchmark datasets

Dataset generator	Benchmark dataset	Year established
DARPA	KDD	1999
UNM [18]	UNM	2004
Creech & Hu, [6]	ADFA-LD	2013

Therefore, ADFA-LD can now be used by the cyber security, machine learning, data mining and soft computing research communities to evaluate the performance of newly proposed intrusion detection systems or algorithms. The new ADFA-LD dataset will continue to be relevant to modern computer technology until this has changed so much that another more advanced new dataset is needed.

4. Evaluation of Intrusion Detection Systems using the ADFA-LD Benchmark Dataset

The new ADFA-LD cyber security benchmark dataset has started to attract the attention of the cyber security research community; for example, Xie *et al*. [21] describe how a one-class support vector machine was used on ADFA-LD to detect intrusion. It was found that the Adduser and Meterpreter were relatively easier to be detected than the Hydra-FTP and Hydra-SSH. It was concluded that the one-class support vector machine was not robust against all kinds of cyber attack. Xie and Hu [22] attempted to extract information from ADFA-LD for the development of a new host-based anomaly detection system using ADFA-LD. The features analysed in their study include the length, common patterns and frequencies of system call traces. It was found that there is an acceptable level of performance with some types of attack. However, the complex behaviour of the modern computer system was not fully understood. Xie *et al*. [21] used K-nearest neighbour and k-means clustering based on the ADFA-LD to explore the potential of reducing the dimensionality of frequency vectors and the optimal distance function was identified.

5. Limitations of the ADFA-LD Benchmark Dataset and Suggestion for Improvement

It is difficult to understand all the attributes of the ADFA-LD cyber security dataset because it is not well described in a way that it can be easily understood by other researchers. A clear description of the input and output attributes of the datasets is critical in the design of data driven intrusion detection systems. To attract the machine learning and data mining community to actively use the ADFA-LD dataset, there is a need for the generators of the dataset to clearly specify input and output attributes. The columns and rows of the dataset are not fully described. These attributes of ADFA-LD should be described in a similar way to the UCI Machine Learning Repository benchmark dataset [23]. This would attract unprecedented attention from the research community. Data preparation and engineering account for 80% of the data mining process [24]. If the datasets are not well understood, meaningful progress might not be achieved in designing data driven intrusion detection systems using the ADFA-LD cyber security dataset. Many potential users of the dataset for evaluation of proposed intrusion detection systems might be discouraged, despite its relevance to present computer technology. Like the UCI Machine Learning Repository, the ADFA-LD datasets should be designed with clear columns and rows in different file formats.

6. Conclusions and future works

This paper reviews the advances made in the cyber security benchmark datasets for the evaluation of machine learning and data mining based intrusion detection systems. The review indicated that the KDD and the UNM datasets, on which researchers relied heavily, have lost their relevance because of the significant changes in computer technology. These changes have also triggered changes in the pattern of cyber crimes. This motivated the generation of the ADFA-LD cyber security benchmark dataset that can handle advances in technology. It has started to attract the attention of the research community, with a number of studies now focusing on its use. The studies in the area of cyber security are expected to continue into the future because the optimal solution to cyber crimes has not yet been achieved. In the future, we intend to create an intrusion detection system based on a hybrid of the cuckoo search algorithm and neural network to be evaluated using ADFA-LD.

Acknowledgement

This research is sponsored by Kulliyyah of Informationa and Communication Technology (KICT), International Islamic University Malaysia (IIUM).

References

1. Joycee, KAM, Parkavi R, Senthikumari R. Network Intrusion Detection & Prevention. *Automat Auton Syst* 2014; **5**:244-245.
2. Chiroma H, Abdulhamid SM., Gital YA, Usman AM, Maigari TU. Academic community cyber caf´es - A Perpetration point for cyber crimes in Nigeria. *Int J Inf Sci Comp Eng* 2011; **2**:7–13.
3. Abdulhamid SM, Haruna C, Abubakar A. Cybercrimes and the Nigerian Academic Institution Networks. *IUP J Inf Tech* 2011; **7**: 47-57.
4. Longe OB, Chiemeke SC. Cybercrime and criminality in nigeria -what roles are internet access points in playing?. *Eur J of Social Sci* 2008; **6**:132–139.
5. Metasploit Penetration Testing Software, http://www.metasploit.com, Accessed November 24, 2014.
6. Creech G, Hu J. Generation of a new IDS test dataset: Time to retire the KDD collection. In: 2013 IEEE Conference on Wireless Communications and Networking (WCNC), 2013; p. 4487-4492.
7. Hu J. Host-based anomaly intrusion detection, In: Handbook of Information and Communication Security, P. Stavroulakis and M. Stamp, editors. Springer Berlin Heidelberg, 2010; p. 235–255.
8. Murtaza SS, Khreich W, Hamou-Lhadj A, Couture M. A Host-based Anomaly Detection Approach by Representing System Calls as States of Kernel Modules. In: 24th International Symposium on Software Reliability Engineering, Pasadena, 2013, p.431-440.
9. Lin YD, Lai YC, Ho CY, Tai WH. Creditability-based weighted voting for reducing false positives and negatives in intrusion detection. *Comput Se* 2013; **39**:460-474.
10. Sinclair C, Pierce L, Matzner. An application of machine learning to network intrusion detection," In: Computer Security Applications Conference, 1999.(ACSAC'99) Proceedings. 15th Annual, 1999; p. 371-377.
11. Chiroma H, Abdul-Kareem S, Abubakar A. A Framework for Selecting the Optimal Technique Suitable for Application in a Data Mining Task. In: Park J.J *et al*. Future Information Technology, 2014; **276**: p.163-169.
12. Julisch K. Data mining for intrusion detection. In: Barbara D, Jajodia S, editors. Applications of data mining in computer security, 2002; p. 33-62.
13. Moradi M, Zulkernine M. A neural network based system for intrusion detection and classification of attacks. In: Proceedings of the 2004 IEEE international conference on advances in intelligent system-theory and applications, 2004; p. 142-152.
14. Kim DS, Nguyen HN, Park J. Genetic algorithm to improve SVM based network intrusion detection system, AINA 2005. In: 19th International Conference on Advanced Information Networking and Applications, 2005; 2. p.155-158.
15. Helali RGM. Data mining based network intrusion detection system: A survey, Novel Algorithms and Techniques in Telecommunications and Networking. pp.501-505. *J Inf Tech* 2010; **7**:47-57.
16. Internet crime complaint center (IC3). The 2013 Internet Crime report . [Online]. Available: http://www.ic3.gov/default.aspx. Accessed November 21, 2014.
17. McHugh J. Testing Intrusion Detection System: a critique of the 1998 and 1999 DARPA Intrusion Detection System evaluations as performed by Lincoln Laboratory. *ACM T Inf Syst Se* 2000; **4**: 262–294.
18. Computer Science Department, "University of New Mexico Intrusion Detection Dataset. [Online]. Available: http://www.cs.unm.edu/_immsec/systemcalls.htm
19. Brown C, Cowperthwaite A, Hijazi A, SoMayaji A. Analysis of the 1999 DARPA/Lincoln Laboratory IDS evaluation data with NetADHICT. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009. p. 1–7.
20. School of Engineering and Information Technology, UNSW, Australia. ADFA Linux data set (ADFA-LD) cyber security benchmark dataset. [Online].Available: http://www.cybersecurity.unsw.adfa.edu.au/ADFA%20IDS%20Datasets/

21. Xie M, Hu J, Slay. Evaluating Host-based Anomaly Detection System: Application of The One-class SVM Algorithm to ADFA-LD,". [Online]. Available: http://icnc-fskd.xmu.edu.cn/doc/invitedSession/FSKD_5_Miao.pdf. Accessed, November 5, 2014.
22. Xie M, Hu J. Evaluating host-based anomaly detection system: A preliminary analysis of ADFA-LD, In: 2013 6th International Congress on Image and Signal Processing (CISP), 2013. pp. 1711-1716.
23. UCI Machine Learning Repository . http://archive.ics.uci.edu/ml/
24. Zhang S, Zhang C, Yang Q. Data preparation for data mining. *Appl Artif Int* 2003; **17**:375-381.